THE RIGAKU JOURNAL
VOL. 11 / NO. 1 / 1994

Contributed Papers

A PRACTICAL APPROACH TO DATA COLLECTION USING THE R-AXIS II

JACK TANNER AND KURT L. KRAUSE*

Department of Biochemistry and Biophysical Sciences, University of Houston, Houston, Texas 77204–5934
** also Department of Medicine, Baylor College of Medicine, Houston, Texas 77030*

During the few years since its introduction, the Rigaku R-axis II imaging plate detector system has become an extremely useful tool of the protein crystallographer [1]. With good spatial resolution properties and broad dynamic range it is capable of collecting excellent data [2]. Owners of R-Axis II systems quickly become experts at data collection and processing using their home systems, but many crystallographers are guest users of the R-Axis and must visit other laboratories to collect their data. Although even the most infrequent users of the R-axis II can acquire good data on their first attempt, it has been our experience that with preplanning before data collection, careful monitoring during data collection, and complete postprocessing, even better data can be obtained. We know this because at the University of Houston we have been frequent guest users of the R-axis facilities at Rice University and at the Molecular Structure Corporation due to the generosity of Professor George Phillips and Dr. Jan Troup. Below we outline our approach to collecting and processing data on the R-axis II with emphasis on the practical aspects of data collection. These comments will be more helpful to the less frequent user of the R-axis II, but hopefully even experienced users will find some useful information here. This approach has been used by us in data collection on a half dozen different proteins from our laboratory, but we certainly do not consider this to be the only way or even the best way to collect data.

Preplanning

Historically, planning a data collection using the oscillation method used to involve reaching a compromise between recording as many full spots as possible on each film while minimizing spot overlap [3, 4]. This approach was driven by the greatly superior merging R values (R_{merge}) obtained on photographic film for full spots relative to partials. With the R-axis II, a well conducted experiment can be expected to yield much lower R_{merge} values than those obtained from film data and closer agreement between the R_{merge} values for full and partial spots [2]. Data collection strategy then shifts somewhat to involve other considerations such as frame exposure time, crystal lifetime, detector distance and the resolution of the planned data set.

Before the Visit

Before visiting an R-axis facility it is advantageous to learn as much as possible about your crystal form and to learn some basics about R-axis data collection. In particular, knowledge of the unit cell, space group, diffraction limits and decay properties of the crystal to be studied will assist in rapidly defining data collection parameters. If you can also obtain information about data collection using the R-axis, you will understand much more of the process as it is going on. The R-axis users manual is an excellent source of information here, as are related publications [1, 5, 6].

Setting up Data Collection

The R-axis II is controlled through a computer interface by a Rigaku software package that is made up of three major divisions, MEASUREMENT MODE, GRAPHIC MODE and DATA PROCESS MODE. These names are indicative of their use since actual collection of images is done in the first mode, viewing images on the terminal screen is done in the second and refinement of parameters and data integration in

the third. Most of the required parameters needed to collect and process data are stored in the crystal.dat file. This information is automatically updated as collection proceeds and shared among the different program modules. Its possible to begin data collection by entering MEASUREMENT MODE, inputting minimal information, and starting image collection. However, we feel that it is advantageous to spend some time studying the behavior of your crystal on the R-axis to obtain information that will help plan more effective and efficient data collection. We start in MEASUREMENT MODE by collecting two twenty minute oscillation images with a phi separation of about 90 degrees. These images will be used to judge diffraction quality and to refine the cell and orientation matrix. Remember to choose a crystal to image plate distance that will supply the desired resolution for your data set.

The next step in planning data collection strategy involves viewing these images in GRAPHIC MODE. Optimizing the contrast and dynamic range of the displayed images is important, and this can be done rapidly through trial- and -error adjustment of the display window. Once optimized the images should be inspected for spot overlap, which might require an adjustment in crystal to image plate distance, and for the quality of diffraction at the edge of the film. Individual high resolution spots can be selected and integrated. If faint spots with low I/σ_I are obtained at the plate edges consider increasing the frame time. Since you will be asked to provide measurement box parameter values for data integration during processing, this is also a good time to view some medium resolution spots under high magnification to get an idea of the best dimensions to use.

Orientation Matrix and Simulation

After making any needed adjustments in exposure time and distance, the orientation matrix can be determined. Following this a data collection simulation can be performed. These operations are done in the DATA PROCESS MODE. Once entered this large software division can be seen to consist of several modules (Fig. 1). Much of the information from MEASUREMENT MODE is passed on in the crystal.dat file, but it is important to inspect these values in SETUP. The orientation matrix is determined and refined in STILL, but auto-indexing and refinement parameters can be input beforehand in SETUP. Either oscillation images or stills can be used successfully for indexing and refinement, and both work equally well in our hands.

Indexing is performed by selecting the frames to be used, picking spots on these frames and selecting auto-index in the still menu. Use the automatic spot picking option, but be sure to set the background level to about −20 in the Frame Information subheading of SETUP before beginning. This will minimize misinterpretation of backscatter off of the beam stop as spots. Only 100 or 200 reflections are needed for auto-indexing, but 200 to 400 should be used for refinement. The default settings in the Auto-Indexing Parameter list usually work, but we have observed failures, especially for cells in which large differences in the lengths of the axes are present. In such a case increasing the number of matching vectors from 3 to 6 and decreasing the angle and length matching error parameters will usually eliminate this problem. Satisfactory auto-indexing has occurred when 95% or more of the spots from the images are accepted.

Following successful auto-indexing, move into the refinement subheading of STILL. There are several refinement parameters from which to choose, but refine only the camera constants and cell parameters skipping spindle, beam parameters and detector two-theta (unless you have a movable stage).

The simulation option appears under OSCILL and allows the user to collect simulated data based on parameters found in the crystal.dat file. In addition to this information the user must supply phi begin, phi end and del phi. Simulation output consists of a frame by frame analysis of the data calculated to appear in the

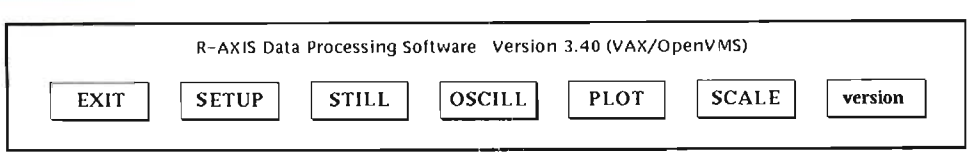

Fig. 1 Main menu of the data processing mode of the R-axis software. The user clicks on these icons to invoke the various program modules.

```
Detector distortion refinement using 148 spots

Starting residual:     0.0162mm    Average C.G.= -0.0011 -0.0001  148 spots
                       ========
Resid. after 1 cyc.:   0.0162mm    Average C.G.=  0.0000  0.0000  148 spots
                       ========
Final RMS residual:    0.0155mm    Average C.G.=  0.0000  0.0000  145 spots
                       ========

    XCEN     YCEN    OMEGA   rDIST     DIST    XSCALE     TILT    TWIST    BULGE
  96.775  100.047   0.0034  1.0013   89.736   0.9996   0.8692   0.2793  -1.7268
   0.002    0.002   0.0021  0.0001            0.0002   0.1005   0.1023   0.2619
```

Fig. 2 An excerpt from a Batch_oscill.log file showing the detector distortion refinement residuals for one frame. The camera constants and their standard deviations are listed at the bottom. Note that the residual decreases as refinement proceeds.

data set. The total number of fulls, partials, overlapping reflections and failures is listed, with failures subdivided into resolution and image plate radii outliers, high Lorentz-polarization reflections and partials found to fall on more than two frames. The del phi can be adjusted and the simulation run again to effect desired change in the number of reflections found in each bin. Some investigators prefer making adjustments to obtain a 2 : 1 ratio of full spots to partial spots, but since the R_{merge} values obtained with partials are close to those obtained with whole spots, we emphasize minimizing spot overlap during the collection while still maintaining a respectable number of wholes per frame. The first listing is followed by a listing of cumulative completeness as a function of frame number. This allows you to determine if the oscillation range you plan to collect will be sufficiently complete before you have committed 1 or 2 days to the collection.

At this point the user has obtained good values for the crystal to image plate distance, frame exposure time, oscillation span per frame and total oscillation range. Data collection can now proceed, but before beginning a few final parameters should be checked. First, set the beam stop trap at zero. This avoids losing some low resolution data. Next verify that the resolution cutoff included in SETUP is correct. Finally increase the maximum detector radius value to about 144 mm. This value will allow for the inclusion of the entire plate in the integration process, if the resolution cut-off is set high enough. Failure to adjust these parameters can result in loss of considerable data in the corners of the plate.

During Data Collection

During data collection the visiting crystallographer may be tempted to assume a more passive role as images accumulate, but close observation during this period can help identify problems during the collection. To do this, begin processing the frames in parallel with image collection. This chase process will integrate the frames and display refinement information as the images are collected. In the output of this process we monitor camera constant refinement, profile shape and data statistics.

Camera constant refinement can often identify slippage between frames and is monitored by looking at detector distortion refinement residuals in the Batch_oscill.log file as each frame is processed (Fig. 2). With a well-behaved crystal, that is, no slippage between frames, these values should be very stable from frame to frame and the refined values should usually stay well below 0.03 mm. If the initial detector distortion residuals are high but rapidly refine to acceptable values, you may be experiencing some slippage and may need to consider using SLIP CHECK in postprocessing (see below). Erratic behavior of these numbers or failure to refine is, of course, indicative of serious problems in the data collection, such as slippage, bad profiles or decay. We also look at the values for XCEN, YCEN, TWIST etc. to see if they have stable, consistent values and verify that they have low residuals.

The next items to view in Batch_oscill.log are the reflection profiles (Fig. 3). Initial box parameters values were set in the Data Collection Parameters subheading in SETUP. By viewing profile output now you can determine if these values were correct, and make any needed changes. The size of the measurement box used in the integration of spot intensities is defined by the parameters NXS, NYS, NCR, NRX, and NRY. NXS and NYS are the width and height of the box in pixels. The other parameters define which pixels are assigned to the peak and back-

standard profile for region 1

SPOTS	AV.INT	AV.X	AV.Y	RMS.DIF	RMS-BG
115.	1267.	3.434	0.768	7.54	0.13

```
0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+
0+  0+  0+  0*  0*  0*  0*  0*  0*  0*  0+  0+  0+
0+  0+  0*  1.  1.  3.  4.  4.  2.  1.  0*  0+  0+
0+  0*  0.  2.  7. 15. 23. 23. 13.  4.  1.  0*  0+
0+  0*  2.  8. 29. 63. 92. 86. 46. 14.  3.  1*  0+
0+  1*  5. 21. 67.138.197.187.108. 34.  6.  1*  0+
0+  1*  7. 33. 99.192.255.225.121. 36.  7.  1*  0+
0+  2*  7. 26. 72.136.178.158. 89. 29.  7.  1*  0+
0+  1*  4. 12. 31. 56. 72. 61. 34. 13.  3.  1*  0+
0+  1*  1.  4.  8. 13. 16. 15.  9.  4.  1.  0*  0+
0+  0+  0*  1.  2.  3.  3.  3.  2.  1.  0*  0+  0+
0+  0+  0+  0*  1*  1*  1*  1*  0*  0*  0+  0+  0+
0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+  0+
```

Fig. 3 An excerpt from a Batch_oscill.log file showing an acceptable reflection profile. Note that the peak is fully contained in the box and the background pixels (denoted by "+") contain zeroes. The boundary pixels (denoted by "*") are not used in the calculation. The measurement box parameters are NXS=NYS=13, NCR=4, NRX=NRY=1.

	FULL REFLECTIONS					PARTIAL REFLECTIONS				
	before LP			after LP		before LP			after LP	
d (A)	NUM	INTEN	SIG	INTEN	SIG	NUM	INTEN	SIG	INTEN	SIG
18.00	1.	45.	8.	1.	0.	1.	-23.	15.	-1.	0.
9.00	3.	1086.	74.	168.	11.	7.	1075.	107.	112.	11.
6.00	21.	715.	40.	124.	7.	14.	1544.	116.	167.	14.
4.50	44.	725.	46.	169.	10.	16.	316.	41.	89.	11.
3.60	69.	867.	45.	252.	13.	44.	625.	47.	150.	12.
3.00	82.	355.	19.	138.	7.	66.	197.	16.	54.	5.
2.57	130.	137.	9.	61.	4.	87.	104.	10.	33.	3.
2.25	171.	59.	6.	33.	3.	119.	67.	7.	19.	3.
2.00	199.	32.	4.	19.	3.	140.	17.	4.	9.	2.
1.80	198.	12.	3.	9.	2.	95.	8.	3.	4.	2.
1.64	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
	918.	191.	12.	64.	5.	589.	161.	15.	36.	4.

Fig. 4 An excerpt from a Batch_oscill.log file listing the number of reflections, the average intensity, and the average standard deviation as a function of resolution. Note that the ratio of the intensity to its standard deviation is greater than 1.5 in all cases.

ground areas of the box. The outer NRY rows and NRX columns of the box are part of the background. NCR defines the corners of the box, which are part of the background. A buffer zone one pixel wide separates the background and the peak. Pixels in the buffer zone are neither part of the peak nor part of the background, and are not used in the calculations. An averaged reflection profile from one of our data processing calculations is shown in Fig. 3. The parameters of this profile are NXS=NYS=13, NCR=4, NRX=NRY=1. Background pixels are denoted by "+" and buffer zone pixels are denoted by "*". All other pixels are part of the peak. The measurement box parameters should be adjusted so that the entire peak fits in the peak area of the box and the background pixels contain values close to zero.

In addition to these two items, we also carefully study the I/σ_I and R_{sym} information located at the end of each frame integration (Fig. 4). This information is printed in tabular form as a function of resolution. We like to maintain I/σ_I greater than 1.5 for the high resolution data, and to observe low R_{sym} values for the reflections with multiple observations per frame. Declining I/σ_I ratios at high resolution can be an early indication of crystal decay or slippage. After the entire collection is complete, the I/σ_I and R_{sym} information can be studied frame by frame to determine if any poor data should be removed from the final processing.

Another direct way of monitoring diffraction during the course of data collection is to view

selected images in GRAPHIC MODE. We inspect these images frequently during collection and usually survey the high resolution areas of the image for adequate diffraction; other indications of crystal decay such as an increase in diffuse scattering are also noted.

Postprocessing

Description of Data Processing Procedures

In this section we discuss the various schemes we use for processing data collected on the R-axis II. These schemes are represented in the flow chart in Fig. 5. The basic processing procedure consists of extracting reflection intensity data from oscillation frames (integrate), merging these data into a single file (merge), performing interframe scaling using an inverse scale factor and a temperature factor (scale), and averaging symmetry equivalent reflection intensities (average). For some proteins the data may be satisfactorily processed using the basic procedure of integrate-merge-scale-average. However, it is our experience that most data sets require additional processing using the post-refinement and slip check options.

In post-refinement the crystal mosaicity, unit cell parameters, and orientation matrix are optimized in order to minimize

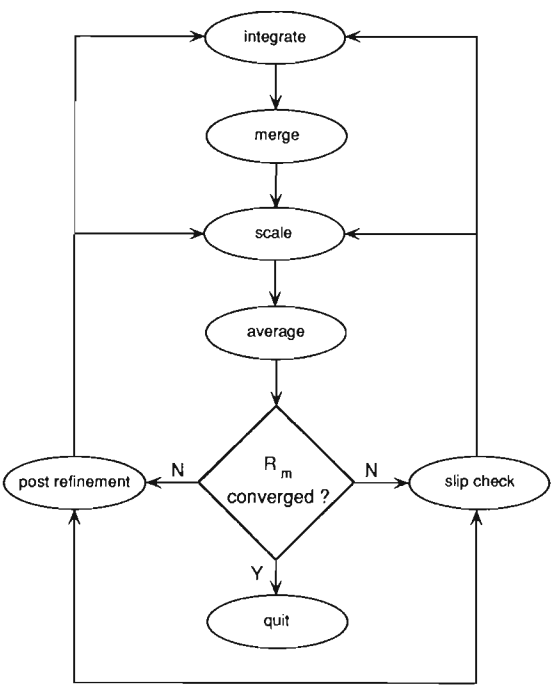

Fig. 5 Flow chart indicating the pathways available for processing data with the R-axis II software. The ovals indicate program modules.

$$\sum |I_{partial} - P_{cal} I_{full}|^2 ,$$

where $I_{partial}$ is the partial reflection intensity, I_{full} is the full reflection intensity, and P_{cal} is the partial reflection factor [5–9]. Proper treatment of partial reflections is crucial in R-axis data processing, thus post-refinement can be a powerful tool.

Slip check, as the name implies, is helpful if the crystal has slipped during data collection. As the crystal slips, the orientation matrix changes, which in turn affects the indexing and integration of reflections. In slip check, an orientation matrix is calculated for each frame. The average missetting angles and the root mean square deviations of the missetting angles from their averages are calculated. From these data one can determine the extent to which the crystal has slipped, and which frames are most affected by the slippage. The software will graphically display the variation of missetting angles as a function of frame, helping the user decide if significant slippage has occurred. The most powerful use of slip check is to use the frame by frame orientation matrices during subsequent reintegration of the images. This is accomplished by selecting the "use of slip orientation" option within the Data Collection Parameter subheading of SETUP prior to reintegration.

The flow chart in Fig. 5 illustrates the various ways to combine postprocessing and slip check with the basic processing procedure of integrate-merge-scale-average. Although we have not exhaustively tested all the possible data processing routes implied by Fig. 5, we have found the following procedures to be useful.

The first calculation we do is the simple integrate-merge-scale-average process, which gives a number of parameters that will allow comparison to other processing runs, such as the merging *R*-factor, scaling *R*-factor, and number of observations. We next invoke post-refinement, followed by scaling and averaging. If the merging *R*-factor decreases, then post-refinement is probably beneficial. At this point, we use the parameters updated in post-refinement to reintegrate the frame data, merge, scale, and average. The resulting *R* values can be compared to the previous *R*-factors to determine whether the reintegration improved the data set. I/σ_I values, as a function of resolution, can often be seen to improve. This procedure may be iterated until the merging *R*-factor, or some other measure of the quality of the data set, converges.

Another approach is to use slip check in the

place of post-refinement in the procedure described in the preceding paragraph. However, this scheme has the disadvantage that the crystal mosaicity is not refined. An alternate method, which allows both refinement of the mosaicity and calculation of the orientation matrix on a frame by frame basis, is to run slip check after post-refinement. This procedure thus consists of the basic integrate-merge-scale-average processing, followed by post-refinement, then slip check, scaling and averaging. If the *R*-factors decrease, the frame data is reintegrated using the parameters updated in post-refinement and the frame by frame missetting angles from slip check. The entire procedure may be iterated until a satisfactory R-factor is obtained.

Application of Data Processing Procedures

We applied the procedures described above to process data from crystals of FMN reductase from *Vibrio harveyi*. FMN reductase is an interesting protein of molecular weight 26,300 that provides reduced flavin to the light emitting enzyme, luciferase by reducing FMN to $FMNH_2$ [10]. The FMN reductase crystals belong to space group $P2_1$ with unit cell parameters a=51.2 Å, b=85.9 Å, c=58.1 Å, β=109.3°, and diffract to beyond 1.8 Å resolution. During data collection the crystal to detector distance was 90 mm and theta was set to zero. The data set consisted of 110 frames with oscillation angle of 1.5° and exposure time of 20 minutes.

Data processing results for FMN reductase are presented in Fig. 6. The seven calculations included in Fig. 6 are labeled A–G. The basic processing scheme, A, gave a data set consisting of 105020 observations with merging *R*-factor, R_m=5.89, and scaling *R*-factor, R_s=8.20. R_m (partials) and R_m (fulls) for this initial processing were 6.19 and 5.65, respectively.

Application of post-refinement without reintegration of the frame data, B, resulted in lower *R*-factors, but 4545 fewer observations. The decrease in *R*-factors observed in B suggested that post-refinement was beneficial, therefore, in calculation C we reintegrated the frames using the parameters optimized in post-refinement. This reintegration resulted in still lower *R*-factors, with R_m=4.46 and R_s=4.44. In addition, the number of observations increased to 117838. Another iteration of the post-refinement followed by integration, D, gave results comparable with those of C, with R_m=4.44, R_s=4.42, 117682 observations, thus the post-refinement processing pathway had converged.

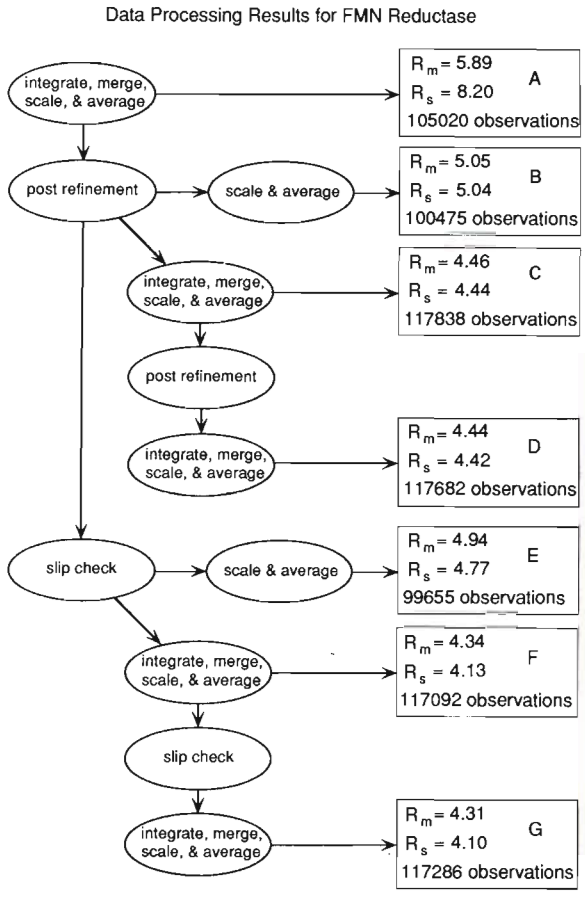

Fig. 6 Results of data processing for FMN reductase. The results of seven calculations are listed and they are labeled A–G. R_m and R_s are the *R*-factors output by AVERAGE and SCALE, respectively. Note that best results are obtained when post-refinement or slip check is combined with reintegration of the frames.

Slip check was applied after post-refinement in calculations E–G. Calculation E, which did not involve reintegration, produced lower *R* values than the basic scheme (A). However, the number of observations was significantly lower than in calculation A. Reintegration, calculation F, produced even lower *R*-factors, R_m=4.34, R_s=4.13, while increasing the number of observations to 117092. Another round of slip check followed by reintegration, G, suggested that the slip check pathway had converged. R_m (partials) and R_m (fulls) for this final processing were 4.86 and 4.15, respectively.

A number of observations can be made from these calculations. First, the basic processing scheme of integrate-merge-scale-average was not sufficient, but application of post-refinement or slip check or both resulted in signifi-

cantly better data sets. Second, the use of post-refinement or slip check without reintegration of the frame data produced lower *R*-factors but the number of observations was also significantly lowered. Upon reintegration using the post-refinement and slip check data, however, even better data sets were obtained in terms of *R* values, and the number of observations. Finally, the iterative schemes of post-refinement or slip check or both, followed by reintegration, converged after two cycles implying that the second cycle is probably not needed.

Summary and Conclusions

We have outlined above our method of planning, collecting and processing data sets on the R-axis II. Careful planning before data collection involves conducting a simulation and inspection of a few frames. The simulation helps in the choice of oscillation angle, angular span of the data collection, and the detector distance. Frame inspection is used to identify overlapping reflections and to gauge the extent of diffraction. Given an estimate of the lifetime of the crystal in the x-ray beam, a collection strategy can be chosen that will give the desired resolution and completeness.

During data collection the refinement of the camera constants and the trend of I/σ_I across the image can provide important clues about the progress of the collection and the quality of the data. Poor refinement of the camera constants can be indicative of crystal slippage. The ratio of I to σ_I can be used to gauge crystal decay and to determine the resolution to include in the postprocessing. Also, if the I/σ_I drops to an unacceptable value, the data collection can be stopped in favor of a new crystal.

Finally, we have also shown that by using a variety of postprocessing schemes we were able to improve the quality of our data. We found that successful processing usually involves several integrations of the frame data using post-refinement or slip check or both. There is probably not one single postprocessing scheme that is ideal for all cases. Rather, the optimal scheme will vary from data set to data set. Therefore, our general approach is to reprocess the data in various ways and compare results. This approach is practical, given the speed of computers and the disk space available in the typical protein crystallography laboratory.

Acknowledgments

We acknowledge the assistance and advice of George Phillips, Fan Yang, Frank Whitby, and Mike Quillin during our data collections at Rice University. We also wish to thank Joe Ferrara and Paul Swepston at Molecular Structure Corporation for their support and advice on data collection. Finally, we thank Mitch Miller for help in preparing this manuscript. This work was supported by grants from the W. M. Keck Foundation, the Welch Foundation, and the State of Texas.

References

[1] Shibata, A., *The Rigaku Journal*, **7**, 28–32 (1990).

[2] Krause, K. L. and George N. Phillips, Jr., *J. Appl. Cryst.*, **25**, 146–154 (1992).

[3] Wonacott, A. J., in *The Rotation Method in Crystallography*, U. W. Arndt and A. J. Wonacott, Eds. (North Holland, Amsterdam, 1977), pp. 75–103.

[4] Wilson, K. and Yeates, D., *Acta Cryst.*, **A35**, 146–157 (1979).

[5] Rigaku, Rigaku Corp., Instruction manuals for R-AXIS IIC and software: Manual numbers ME200KA1, ME201LP1, ME201LS1.

[6] Higashi, T., Rigaku Corp, PROCESS: A Program for Indexing and Processing R-AXIS II Imaging Plate Data (1990).

[7] Schutt, C. E. and Winkler, F. K., in *The Rotation Method in Crystallography*, U. W. Arndt and A. J. Wonacott, Eds. (North Holland, Amsterdam, 1977), pp. 173–186.

[8] Schutt, C. E., Dissertation, Harvard Univ. (1976).

[9] Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S., and Tsukihara, T., *J. Appl. Cryst.*, **12**, 570–581 (1979).

[10] Lei, B. and Tu, S.-C., in *Flavins and Flavoproteins*, K. Yagi, Eds. (1994), in press.